The Universal Patient Identifier: A Discussion and Proposal

Paul C. Carpenter, M.D., Christopher G. Chute, M.D., Dr.P.H.
Division of Endocrinology and Section of Medical Information Resources,
Mayo Clinic, Rochester, MN

The broad-based need for a standard, robust patient identifier is reviewed. Advantages and difficulties of the Social Security Number are considered. An alternative number, composed of birthdate, location, and check digits, is proposed. Logistic extensions of the geographically based schema are made for provider and facility identification. Several variations of representation are described in base 34, which make the composite number short, easily derived and decoded, and robust.

INTRODUCTION

Electronic medical records are the digital repository of patient experiences in health and disease. Following a person over time allows observation of a developing disease, interventions and their efficacy, and the eventual outcome of a person's success in disease management. Analyzing the longitudinal experiences of many enables epidemiologic inference about the relative merits of alternative treatments and interventions. Fundamental to this scenario is the ability to observe the fate of patients as they migrate from provider to provider and place to place.

The specter of a universal patient identifier has never failed to provoke controversy. Biblical reference to the "mark of the beast" (666) in Revelation are perhaps the earliest, but by no means the last reference to religious, moral, or psychological objections to human enumeration. A landmark 1973 Department of Health and Human Services report rejected any move toward "Standard Universal Identifiers" on confidentiality grounds. Nevertheless, the opportunity to study health practice and generate data-driven clinical guidelines has sapped support for the concern. Gigabytes of Medicare data, linked by a scrambled but still consistent patient identifier, are now the daily target of health service researchers nationally.

The Social Security Number (SSN) is the *de facto* linkage within and between governmental agencies. Significantly, it is also the major linkage in the burgeoning credit bureau industry, where the fortunes and fate of millions are viewed in the best traditions of free enterprise. Numerous other examples of the non-healthcare/pension use of the SSN for identification develop constantly. In the near future this means of patient identification will likely persist and expand outside the federally managed health systems. This is driven by the broad (but incomplete) distribution of SSN in our population and the ease of implementation of an available, though flawed, enumerator.

Confronted with the confidential linkage problems to existing data bases, the lack of a check digit, a forty-year space of unverified assignment, and a huge pool of unvalidated numbers, many in the health care community have little enthusiasm for embracing the SSN as a universal patient identifier, except as a stop-gap device. As we become more closely allied with our global village, extension beyond the United States poses additional problems. While the use of SSN, hopefully with a check-digit addition, may serve this country's short-term needs, it is time to rationally consider alternative truly Universal Patient Identification (UPI) systems.

UNIVERSAL PATIENT IDENTIFIER

Desirable features of a patient code include uniqueness, verifiability, entry reliability, and practical administrative assignment and tracking. Additionally, it should be based on immutable person properties rather than those which may change by political or personal whim, such as town, state, country, or last name.

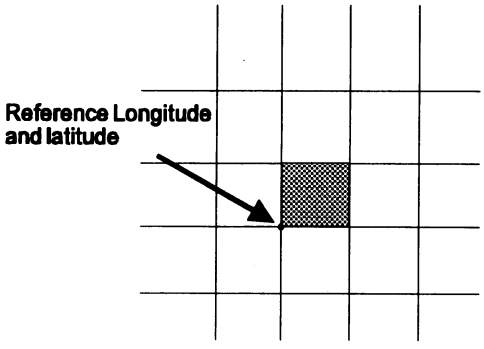
0195-4210/92/$5.00

Our proposed model for the UPI consist of a series of three universal and immutable values plus a check digit: 1) A 7 digit date code, 2) A 6 digit geographic code that relates to the place of birth or entry into the healthcare system, 3) A 5 digit sequence code to identify people born on the same date in the same geographic area, plus 4) A single check digit. The model is distinct from its final appearance, which is modified in the next section to avoid a 19 digit monster.

The 7 digit date field allows for 1000 years by containing the last three digits of the year; 2 digits each for month and day are also allocated. If this were only 6 digits it would allow for only 100 years, which poses significant problems for longitudinal study of unique persons in population cohorts. Further, we are optimistic enough to believe healthcare will improve to the point where many more of us lives beyond 100 years.

The 6 digit geographic code identifies world grid coordinates using 360 degrees for longitude and 180 degrees for latitude in the range as opposed to 180° east/west or 90° north/south. Reference 0 degrees are the north pole and Greenwich. These coordinates would indicate the region of birth, or if that is not known, the first point of entry into the healthcare system. This avoids the problem of a country identifier and the fact that country names and boundaries are subject to change (the former Yugoslavia and Soviet Union serving ample example.) Geographic coordinates are not subject to change and are universally recognized.

One degree increments represent a maximum area of only 70 miles square (at the equator), thus the UPI could be administered locally. Convention would dictate the area of jurisdiction to be that north and east of the grid coordinates, bound by the next increment latitude and longitude lines. This is shown as:



For example, a person born on March 1, 1993, in Minneapolis, MN the code has the appearance: 9930301^044273^00047^2.

Within this area a birth occurring in a hospital, home or even a taxi could be linked to a local jurisdiction that would administer the registration of the UPI. The local organization would then forward the code to an international registry administered by a body such as the World Health Organization (WHO).

If geographic coordinates overlapped areas of political or administrative jurisdiction, the sequence number is large enough to be broken into ranges such as sub-area #1 - 1 to 10000, #2 - 10001 to 20000, etc. and would be divided up according to a locally determined convention.

It would obviously be difficult for an individual to recall their 19-digit sequence on request. Transformation of the sequence to a shorter representation has appeal, and is considered in the next section. The check digit is also discussed in the next section as it pertains to the concise representation of the UPI. However, the principles of prime-modulo computation can apply to any arbitrary designation.

One of the issues that must be dealt with is the entry of an individual into the healthcare facility who does not have the UPI in their possession, has never had one issued, or is incapable of relating their number to caregivers (unconscious, etc.). This is handled simply by assigning a temporary UPI, distinguished by the addition of a leading "T." This would indicate that the UPI does not link back to a previously established record. If additional demographic data is acquired in the registry process, the central registry at an organization like WHO would be able to compare records, and if matches were found, link this back to multiple records and update them under the proper UPI.

The same UPI system could be used as a Universal Provider Identification. It is presumed that all healthcare providers will have been issued a UPI. For purposes of healthcare provider identification their personal UPI could be modified by adding an initial digit, such as "P"(rovider). Other characters could be used to identify provider types, e.g., MD, RN, LPN, PA, etc., if necessary.

Similarly, a Universal Site Identification for medical care location could use the same structure simply deleting the date digit section, e.g. ^046267^00021^2123^. In this case the five digit sequence code would identify the specific hospital, clinic, or office in the geographic area as site of care.

A perplexing problem is the specific linking of an individual with their UPI. Dishonesty or other motives may prompt an individual to fabricate identifying information and thereby be issued a new and superfluous UPI or to be indexed on the UPI of another individual. In the future, positive identification may have to involve automated finger print recognition or comparable technology, definitively linked to a UPI.

We feel this approach for the UPI is reasonable and flexible in terms of its administration. It can be used to retroactively update existing records based on known information from the birth record. It is well suited to be administered by an existing infrastructure of current governing agencies around the world. Issues will still remain in terms of patient/provider confidentiality, but these concerns can be worked around the UPI and are independent of the model or number format.

CONCISE REPRESENTATION ALTERNATIVES

The Universal Patient Identifier as outlined may be regarded as a 19 decimal digit number string. The components are detailed in Table 1. The purpose of this section is to consider how similar data might be more succinctly represented.

Table 1
Decimal-Based Model of UPI

Data	Decimal Digits
Year of birth (0-99)	3
Month of birth	2
Day of birth	2
Latitude (0-180)	3
Longitude (0-360)	3
Strata sequence	5
Check digit	1

Base Considerations

The numeric base in which a number is represented can extraordinarily influence its physical length. Let us observe by example the length of an arbitrary decimal number (1234567890) in various numeric bases, depicted in Table 2.

Table 2
Representation of 1234567890_{10} in Other Bases

Base	Number
2:	1001001100101100000001011010010
8:	11145401322
10:	1234567890
12:	2a5555016
16:	499602d2
34:	t5uraa
36:	kf12oi
58:	1T5u1g
62:	1ly7vk

All these "numbers" are printed using the digits: 0-9, a-z, and A-Z. The major point is that the higher based numbers are shorter than the decimal or lower base numbers. Examining the base 36 representation shows that the letter "o" might be confused with zero, similarly the letter "l" in the base 62 number might be confused with one. For this reason, bases 34 and 58 eliminate the letters "o" and "L" as allowable digits.

Considerable shortening occurs by base 34. Also, the higher bases (beyond 36) rely on significant upper and lower case digit distinction, a practice highly unlikely to survive many manual transcriptions. We propose that base 34, which disallows the letters "o" and "L" for clarity, be adopted as the numerical base for the UPI.

Bit Minimalism

Let us attempt to represent the three major UPI components, geo-code, date, and sequence, with a minimum number of bits for each (check digits will be dealt with separately). Latitude and

longitude have 180 and 360 potential values respectively, thus the geo-code has 360 x 180^2 combinations (64,800) which fit well in a 16 bit number (decimal range 65,536). To maximally compress date, let us count the number of days since a date picked to be before anybody living could be born, e.g. 1880. Allowing 16 bits for the date code gives about 180 years of date information, making this UPI version useful until 2059. Twelve bits (4,096) for sequence numbers within a day totals our bits to 44 (decimal range: 17,592,186,044,416). Such information can fit in a 8 digit base 34 number. These are minimalist assumptions. Adding two additional base 34 digits allows us to fit a 50 bit binary number easily. Table 3 summarizes these options and ranges.

Table 3
Base 34 Number

	8 Digit		9 Digit	
	bits	range	bits	range
Geo-code	16	360 x 180	16	360 x 180
Date	16	1880-2059	18	1880-2600
Sequence	12	4,096	16	65,536
Total bits	44		50	

Clearly, if bit manipulations are adopted, the 10 digit number will afford a much longer useful life through the year 2600. Alternatively, one could borrow bits from the sequence number in the 8 digit number, and extend the longevity of a 8 digit base 34 UPI.

If the issuing of sequence numbers are limited to assignment on birth date (as opposed to a disaster enrollment date), how many numbers might be needed? Let us assume that present day India, a crowded nation, sustained an unlikely 4% birth rate. Knowing the nation's area to be 1,269,339 square miles and the population to be about 700,000,000, this computes out to be approximately 259 geo-codes (at 70^2 miles, a conservative estimate at that latitude) and yields an average of 296 births/geo-code/day. Thus, a sequence number of 2^{10} (decimal range: 1024) might be adequate, making the 8 digit base 34 digit more attractive and usable through the year 2600.

Practical Computing

It may be that every issuer of the UPI might have access to a PC level computer, more than adequate to do the bit computations and base changes implied above. Nevertheless, the coding and decoding of bits into base 34 definitely requires a computer. If we allow just one more base 34 digit, we can make the process practical with a pencil, and therefore comprehensible. The premise here is to group base 34 digits into logical, tractable code elements. Table 4 illustrates this model.

Table 4
Concise UPI Model for Pencil and Paper

Data	Base 34 Digits	Range
Latitude	2	180
Longitude	2	360
Day	1	31
Month	1	12
Year	2	1880-3036
Jurisdiction	1	34 39,304
Sequence	2	1,156

In Table 4, we have broken the sequence number into jurisdiction, allowing up to 34 centers, such as might be found in a large metropolitan center, to independently assign over 1000 UPIs each day within a geo-code, without the risk of conflict. However, if a great premium is given to a short number, the jurisdiction digit could be dropped altogether, and yield a practical 10 digit UPI. Sufficient sequence range remains for birth assignment, as outlined by the India birth example.

This format can be issued without a computer because:

- The four digit geo-code is fixed for a given center.
- Day and month are easily converted to base 34 digits.
- Year changes slowly, and will be known (retrospective assignment can be computed).
- Jurisdiction is fixed.
- Sequence number can be crossed off a printed sheet for each day.

Back computing this 11 digit format can be done easily on the back of an envelope. It also has the desirable property of being less likely to spell vulgar words by chance in any language, since at least 3 digits never exceed the value of "B".

The Check Digit

This problem can be considered independently from the UPI. Its length will be driven by our insecurity, one base 34 digit given a 1/34 chance of error, two digits protected up to 34^2 (1,156), and so on. Our comfort level rests with a single check digit, which would neatly round off to 12 total digits in the UPI.

Classic check digit algorithms are designed to protect against mis-keys and digit inversions. The most basic algorithm which does both is:

- multiply each digit by its corresponding prime
- sum the products
- modulus the sum by the check precision

Consider the 5 digit number 12345:

digit		prime	product
1	x	1	= 1
2	x	3	= 6
3	x	5	= 15
4	x	7	= 28
5	x	11	= +55
			105 mod 34 = 3

This algorithm can accommodate any number of digits in the UPI, and yield any number of digits as the check. For example, if two base 34 digits are desired in the check, then the modulus is by 34^2 (1,156). Computation of the check digit should be done by computer at the issuing center, or special purpose pocket calculators.

SUMMARY

We have outlined above a proposal for the adoption of a UPI which is based on immutable descriptors, not subject to the whims of personal identification or politics. Additionally, we suggest the adoption of a base 34 character representation of the UPI for personal memory and ease of entry into the electronic medical records of the future. Use of this UPI allows the bridging of political boundaries without infringement on local custom or existing systems.